Meteorologische Zeitschrift, Vol. 25, No. 6, 721–738 (published online July 16, 2015)

MiKlip

Probabilistic evaluation of decadal prediction skill regarding Northern Hemisphere winter storms

Tim Kruschke[1,2], Henning W. Rust[1*], Christopher Kadow[1], Wolfgang A. Müller[3], Holger Pohlmann[3], Gregor C. Leckebusch[4] and Uwe Ulbrich[1]

[1]Institute of Meteorology, Freie Universität Berlin, Berlin, Germany
[2]GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany
[3]Max-Planck-Institute for Meteorology, Hamburg, Germany
[4]School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, UK

(Manuscript received August 1, 2014; in revised form March 24, 2015; accepted May 25, 2015)

Abstract

Winter wind storms related to intense extra-tropical cyclones are meteorological extreme events, often with major impacts on economy and human life, especially for Europe and the mid-latitudes. Hence, skillful decadal predictions regarding the frequency of their occurrence would be of great socio-economic value. The present paper extends the study of Kruschke et al. (2014) in several aspects. First, this study is situated in a more impact oriented context by analyzing the frequency of potentially damaging wind storm events instead of targeting at cyclones as general meteorological features which was done by Kruschke et al. (2014). Second, this study incorporates more data sets by analyzing five decadal hindcast experiments – 41 annual (1961–2001) initializations integrated for ten years each – set up with different initialization strategies. However, all experiments are based on the *Max-Planck-Institute Earth System Model* in a low-resolution configuration (MPI-ESM-LR). Differing combinations of these five experiments allow for more robust estimates of predictive skill (due to considerably larger ensemble size) and systematic comparisons of the underlying initialization strategies. Third, the hindcast experiments are corrected for model bias and potential drifts over lead time by means of a novel parametric approach, accounting for non-stationary model drifts. We analyze whether skillful probabilistic three-category forecasts (enhanced, normal or decreased) can be provided regarding winter (ONDJFM) wind storm frequencies over the Northern Hemisphere (NH). Skill is assessed by using climatological probabilities and uninitialized transient simulations as reference forecasts. It is shown that forecasts of average winter wind storm frequencies for winters 2–5 and winters 2–9 are skillful over large parts of the NH. However, most of this skill is associated with external forcing from transient greenhouse gas and aerosol concentrations, already included in the uninitialized simulations. Only over East Asia and the Northwest Pacific, the Northwest Atlantic as well as the Eastern Mediterranean the initialized hindcasts perform significantly better than the uninitialized simulations. While no significant differences are evident between anomaly- and full-field-initialization, initializing the model's ocean component from *GECCO2*-ocean-reanalysis yields slightly better results than from *ORA-S4*, especially over the Northeast Pacific. Additionally, it is shown that the novel parametric drift-correction approach – estimating potential cubic drifts with parameters linearly changing in time – is more appropriate than the standard procedure – estimating constant model drifts via the lead-time-dependent bias – and, hence, yields higher skill estimates.

Keywords: decadal prediction, winter storms, drift-correction, MiKlip

1 Introduction

In addition to externally forced climate change on centennial time scales, climate evolution exhibits substantial variability driven by natural processes. On decadal timescales these variations coincide with typical planning horizons of political and economic stakeholders. Thus, climate predictions for the next decade(s) would be of great socio-economic value (Solomon et al., 2011) if proven to be of significant skill with respect to relevant parameters. Winter storms have been responsible for approx. € 25.2 bn overall losses in Europe during 1980–2006 (Munich RE Group, 2008, in values of 2006). They are the most expensive type of natural catastrophe in this area, primarily because of direct wind damage. In North America winter storms throughout the period of 2002–2011 caused an average overall loss of US$ 1.8 bn per year (Munich RE Group, 2013, in 2011 values). Losses in North America are mainly due to snow amounts and related damages, though. These numbers highlight the potential relevance of decadal predictions in this respect, further emphasized by the fact that the frequency of such events exhibits a high degree of inter-annual to multi-decadal variability (see e.g. Donat et al., 2011; Nissen et al., 2014a; Welker and Martius, 2014). The German initiative *Mittelfristige Klimaprognosen* (MiKlip) is dedicated to the development of a model system to produce skillful predictions for up to a decade ahead (Pohlmann et al., 2013a). The present

*Corresponding author: Henning W. Rust, Institute of Meteorology, Freie Universität Berlin, Carl-Heinrich-Becker-Weg 6–10, 12165 Berlin, Germany, e-mail: henning.rust@met.fu-berlin.de

DOI 10.1127/metz/2015/0641

Gebrüder Borntraeger Science Publishers, Stuttgart, www.borntraeger-cramer.com

study evaluates five MiKlip-experiments conducted so far, focusing on the predictive skill regarding the frequency of winter wind storms over the Northern Hemisphere.

Following some pioneering studies regarding potentials and limitations of decadal climate predictions (see Meehl et al., 2009, for a thorough review), the *Coupled Model Intercomparison Project* in its fifth phase (CMIP5) included a new framework for initialized decadal predictions. This was done to set the scene for a coordinated assessment of current earth system models' ability to produce reliable climate predictions. One of the hindcast experiments analyzed in this study (*baseline0*, see Section 2.1) actually contributes to the CMIP5 decadal experiments.

The basic idea behind decadal predictions is to initialize the models with an observed climate state. It is expected that these simulations subsequently follow to some degree the observed climate evolution, as they should include the unforced component of climate variability (Taylor et al., 2012). Additionally, the initialization might correct potential inaccurate responses to previous external forcing. As opposed to this, standard transient simulations are started from arbitrary states of afore-conducted control experiments representing pre-industrial conditions. Subsequently, they are run throughout the whole historical period using observed external forcing and potentially continued, employing specific forcing scenarios. Hence, only the respective forced component of these simulations is comparable with historical climate evolution. These transient runs are typically started several decades previous to the validation period for decadal predictions and they never experience any "contact" to actual climate states. They are often called *uninitialized* in the context of decadal prediction. However, initializing decadal predictions with observed states is not trivial. Several studies exist, searching for optimal strategies in this respect (e.g. Matei et al., 2012; Smith et al., 2013; Hazeleger et al., 2013; Counillon et al., 2014; Polkova et al., 2014). In general their results are promising but further improvements are necessary to enable the models catching the right phase (and amplitude) of actual climate evolution.

Another issue is the existence of model biases; if these are constant over time, they are rather unproblematic and can be overcome by standard evaluation procedures based on (climatological) bias adjustments or analyzing anomalies. However, biases represent some challenge if not constant over time. While a bias of a single prediction growing over lead time might be just a bad forecast, more systematic issues may exist, potentially masking any predicted climate signal. In this context, it is essential to understand the reasons of such systematically changing biases. One possible reason is the existence of externally forced long-term changes of the model which differs from those evident in the chosen observational data. Another reason is that the initialization (especially full-field initialization) sets the model close to the observational state, which might be incompatible with its own preferred (systematic error) state (Meehl et al., 2014). In that case the model will exhibit systematic drifts back towards its equilibrium over simulation time. For coupled models, including an ocean component with its inertial character, these (not necessarily monotonous) drifts might be of significant influence on the model results for several years or even decades, depending on the order of the initial shock. Such systematic but non-stationary biases require appropriate bias-correction methods (see e.g. Kharin et al., 2012; Hawkins et al., 2014) applied to decadal climate predictions before they can be properly analyzed regarding their predictive signals and the associated skill of these predictions.

In recent years, more and more studies have become available, dedicated to skill assessment of existing individual forecast systems (e.g. Müller et al., 2012; Müller et al., 2014; Boer et al., 2013; Goddard et al., 2013) or multi-model ensembles (e.g. van Oldenborgh et al., 2012; Doblas-Reyes et al., 2013; Meehl and Teng, 2014). Most of these studies consider fields of primary meteorological parameters, such as mean surface air temperature and precipitation. Other studies focus on specific meteorological phenomena or indices, e.g. Garcia-Serrano and Doblas-Reyes (2012), assessing decadal prediction's skill regarding large scale mean temperatures, or Scaife et al. (2014), analyzing predictions of the *Quasi-Biennial Oscillation* (QBO). So far, very few studies deal with the predictive skill with respect to the frequency of meteorological events such as temperature and precipitation extremes (Eade et al., 2012; Hanlon et al., 2013) or the occurence of tropical or extra-tropical cyclones (Smith et al., 2010; Kruschke et al., 2014). As the common understanding of the climate system emphasizes the role of the ocean for decadal climate variability, many studies concentrate on this subsystem, analyzing (global) fields of sea surface temperatures (SST) and upper ocean heat content (e.g. Matei et al., 2012) or the oceanic variability of specific regions and oceanic phenomena such as the North Atlantic and the Atlantic Meridional Overturning Circulation (AMOC; e.g. Kröger et al., 2012; Pohlmann et al., 2013b) or the Atlantic Subpolar Gyre (Yeager et al., 2012; Robson et al., 2012; Robson et al., 2014). Most existing studies on decadal prediction skill follow a deterministic verification approach, comparing the forecast ensemble mean and observations. Only few studies try to incorporate the full ensemble information by employing probabilistic verification approaches (e.g. Goddard et al., 2013; Kruschke et al., 2014; Stolzenberger et al., accepted).

This paper extends the study of Kruschke et al. (2014) which showed that the first two development stages of the MiKlip decadal prediction system exhibit promising levels of skill with respect to the frequency of (intense) Northern Hemisphere extra-tropical cyclones. Instead of addressing these general meteorological features (diagnosed from the Laplacian of sea-level pres-

Table 1: Overview of analyzed hindcast experiments (parenthesized ensemble size available only for every fifth initialization: 1961, 1966, . . . , 2001)

Hindcast experiments	Initialization atmosphere	Initialization ocean	Ensemble members
baseline0	none	anomalies from NCEP/NCAR-forced ocean run	3 (10)
baseline1	full-fields from ERA40/ERA-Int.	anomalies from ORA-S4	10
ORAff	″	full fields from ORA-S4	10
GECCOano	″	anomalies from GECCO2	3
GECCOff	″	full fields from GECCO2	3

sure), the present study analyzes potentially damaging (surface) wind storms (see Section 2.2). This approach essentially means that the focus of this paper is tailored to primary interests of economic and societal stakeholders. Besides this more applied emphasis compared to Kruschke et al. (2014), the present study is more elaborated in terms of adjusting for potential model drifts over forecast lead time (see Section 4). Last but not least, we consider more hindcast experiments (a total number of five instead of two used by Kruschke et al., 2014) as well as different combinations of these experiments (see Section 2.1) in order to derive more robust estimates of predictive skill and systematically compare the performance of the underlying initialization strategies.

The following questions are addressed:

1. Do decadal predictions of winter storm frequency provide significantly more information for the next few years than the climatological forecast (i.e. using the climatology as a forecast)?
2. Does initialization from actual climate states (as realized so far) provide any additional value compared to uninitialized simulations (including responses to external forcing only) and linear approximations of long-term change?
3. Is any of the so-far-realized initialization strategies clearly superior to the others in predicting winter wind storm frequencies?
4. What is the effect of the chosen parametric drift-correction approach on estimations of predictive skill compared to the standard non-parametric procedure, used in many other studies?

Section 2 describes all used data, that is first of all the different decadal hindcasts, but also the uninitialized simulations and reanalyses. Additionally, Section 2 contains the methods to identify winter storm events, the calculation of winter storm frequencies, as well as the chosen probabilistic verification metric. Section 3 is dedicated to the analysis of systematic differences, such as biases and deviating long-term trends, between MPI-ESM-LR and reanalysis. Subsequently, an appropriate approach to statistically adjust the model for these systematic deviations is shown in Section 4. The results of probabilistic hindcast verification are to be found in Section 5, while Section 6 summarizes the paper and its conclusions.

2 Data and methods

2.1 Data

All model simulations analyzed in this study are conducted using the *Max-Planck Institute Earth System Model* in a low-resolution configuration (MPI-ESM-LR, see Giorgetta et al., 2013). The atmospheric component of MPI-ESM-LR is ECHAM6 (see Stevens et al., 2013) in T63L47-resolution (approx. 210 km horizontal grid spacing at the equator) while the ocean is represented by MPIOM with GR15L40-resolution (grid spacing ranging from approx. 15 km around Greenland to 185 km in the tropical Pacific; see Jungclaus et al., 2013).

Five sets of decadal hindcasts are analyzed in this study. Each consists of 41 hindcasts, initialized annually at 1st January 1961–2001 and integrated for ten years each. The five hindcast sets differ with respect to the underlying initialization strategies, as summarized in Table 1. The starting-point within the *MiKlip*-initiative is called the *baseline0*-system, identical to the decadal prediction set up used for the CMIP5-exercise. In order to generate the initial conditions for each *baseline0*-hindcast, an ocean-only-experiment was conducted, forcing MPIOM with atmospheric data from *NCEP/NCAR*-reanalysis (Kalnay et al., 1996). The ocean temperature and salinity anomalies of this experiment are then used for nudging a run of the coupled model. The initial states of the decadal hindcasts are taken from this run. Only the oceanic component (i.e. not the atmosphere) is relaxed in this way to the observed state for *baseline0*. This hindcast experiment consists of three ensemble members for most initializations, while for every fifth hindcast (initialization year 1961, 1966, . . . , 2001) ten ensemble members were produced. *Baseline0* was introduced by Müller et al. (2012), also analyzing its performance with respect to decadal forecasts of seasonal mean surface temperatures.

The four other hindcast sets were initialized by nudging the coupled model towards fields directly derived from reanalyses. For the atmospheric component, this is done identically for all four sets via full-field-initialization from ERA40 (Uppala et al., 2005, for the initializations 1961–1989) and ERA-Interim (Dee et al., 2011, for initializations since 1990). The difference between these four hindcast sets is to be found

Table 2: Overview of analyzed multiple system ensembles, produced by combining different hindcast experiments, listed in Table 1 (parenthesized ensemble size available only for every fifth initialization: 1961, 1966, . . . , 2001)

Multiple system ensembles	Hindcast experiment combinations	Ens. mems.
ORA-Ens.	*baseline1* + *ORAff*	20
GECCO-Ens.	*GECCOano* + *GECCOff*	6
ano-Ens.	*baseline1* + *GECCOano*	13
ff-Ens.	*ORAff* + *GECCOff*	13
Grand Ens.	*baseline0* + *baseline1* + *ORAff* + *GECCOano* + *GECCOff*	29 (36)

in the ocean initialization. Two sets were initialized from the ORA-S4-ocean-reanalyses (Balmaseda et al., 2013) while GECCO2 (Köhl, 2015) was used for the other two. Each of these pairs is split up by using full-field-initialization (i.e. nudging the ocean model to absolute values of the ocean reanalysis) and anomaly-initialization (nudging to values resulting from ocean reanalysis anomalies that were added to the model's climatology), respectively. The system characterized by anomaly-initialization from ORA-S4 is called *baseline1* in the MiKlip-context and was elaborately described by Pohlmann et al. (2013a), studying the decadal prediction skill of this system in comparison to both, *baseline0* and a prediction system with different resolution. The full-field initialized hindcasts based on ORA-S4 are part of the next development stage of the MiKlip prediction system – the *prototype system*. As they do not consitute the full prototype ensemble, we call them simply *ORAff* for the current study. The two GECCO2-initialized systems are accordingly called *GECCOano* and *GECCOff*. All ensembles are generated by lagged-day-initialization.

The five hindcast sets also differ in terms of ensemble size (see Table 1), incorporating three or ten ensemble members. These very small ensembles pose a challenge for probabilistic hindcast verification. For parameters exhibiting low signal-to-noise-ratios deterministic verification (of the ensemble mean) is similarly difficult, though (see Sienz et al., submitted). This means that verification results will suffer from high uncertainty, which – in the field of decadal predictions – is additionally fueled by the limited number of independent hindcasts, i.e. initializations. To reduce this uncertainty stemming from the small ensemble sizes, but also to more systematically examine the effects of the different initialization strategies, we additionally analyze multiple-system ensembles. These are composed by different combinations (see Table 2) of the above-mentioned hindcast experiments after separately adjusting them according to Section 4. Two of these multiple system ensembles comprise all available ensemble members produced by anomaly- (*ano-Ens.*) and full-field-initialization (*ff-Ens.*), respectively. Two more multiple system ensembles compound all members initialized from ORA-S4- (*ORA-Ens.*) and GECCO2-ocean-reanalysis (*GECCO-Ens.*), respectively. Additionally, a *Grand Ensemble* is put together by combining all available hindcasts, including *baseline0*.

The benefit of the initialization must be assessed. As mentioned before, this can be done by computing the skill against a forecast representing climatological conditions. Knowing about observed climate change assigned to transient greenhouse gas and aerosol forcing, the larger challenge of the initialized forecasts is to outperform the so-called uninitialized simulations. An ensemble of ten such simulations is generated by starting them from randomly chosen states of a long pre-industrial coupled control simulation. According to the years covered by the analyzed hindcasts, we examine only those model years representing the period 1961–2011.

To assess the quality of the hindcasts (and the uninitialized runs) we use the reanalyses of the *European Centre for Medium-Range Weather Forecasts* (ECMWF). Winter wind storm frequencies are determined per boreal winter half year (ONDJFM, see below) in this study. Thus, ERA40 is used for the winters 1961/62–1989/90 and ERA-Interim for 1990/91–2011/12 which is in correspondence to the above-mentioned atmospheric initialization of four hindcast experiments.

2.2 Winter storms: identification and frequency calculation

Northern Hemisphere winter wind storms are identified from reanalysis and model data (north of 0 °N, including a sponge-zone between 0 °N and 10 °N) via an objective scheme, based on 6-hourly instantaneous surface wind speeds. The latter are scanned for meso-alpha- to synoptic-scale contiguous areas (minimum of 150.000 km^2) of extreme values exceeding the local climatological 98th percentile. These strongest 2 % of surface winds are often associated with damage (e.g. Klawa and Ulbrich, 2003; Leckebusch et al., 2007; Schwierz et al., 2010). The identified fields of wind speed extremes are tracked over consecutive time steps by applying an iterative algorithm based primarily on a nearest-neighbor search. For each time step, a point-like position of the storm field is calculated, eventually yielding the track of the respective winter storm. Only winter storms tracked for at least 18 hours are considered in this study. This algorithm was first introduced by Leckebusch et al. (2008), has been further developed since then (see Kruschke, 2014, for a thorough description of the actual scheme), and used in several studies on climatologies (e.g. Nissen et al., 2010; Nissen et al., 2014a; Nissen et al., 2014b; Pardowitz et al., submitted), and seasonal prediction skill over the North Atlantic and Europe (Renggli et al., 2011).

In the present study, winter storm frequencies are calculated as track densities on a pre-defined 2.5 °-grid as

the number of tracks crossing a region of 1000 km radius (great circle distance) around the respective grid points per boreal winter half year (ONDJFM). To avoid boundary effects and focus on extra-tropical phenomena, only results north of 30 ° N are used for all further analyses in this study.

To prevent from inconsistencies within the observational reference, the winter storm frequencies of ERA40 were corrected to match mean and variance of ERA-Interim for the 22 winters existent in both data sets (1979/80–2000/01). Additionally, correlations of winter storm frequencies between the two data sets were calculated, based on these 22 winters. Those grid points exhibiting insignificant (p-value > 1 %) correlations were rated as not reliable and thus excluded from all further analyses (masked in gray for all Figures).

2.3 Probabilistic hindcast verification

Winter storm frequencies for individual years were mapped onto one of three categories: below normal, normal, or above normal. The categories are separated by the first and second terciles, classifying a given forecast below normal if it falls below or onto the first tercile and above normal if it is higher than the second tercile. Both thresholds are empirically derived from the 51 winters 1960/1961–2010/2011 (i.e. 17 values in each category for the reanalysis data). For the model, all ten ensemble members of the uninitialized runs and the respective period (i.e. 510 model winters) were used to derive corresponding thresholds. In addition to individual season analyses, the hindcasts are verified for perennial averages. For these temporally aggregated hindcasts the category thresholds are equivalently calculated as empirical terciles based on running means over the mentioned period. The window width for these running means is chosen equivalent to the respective aggregation for hindcast verification, e.g. 4 y-running means used to derive thresholds for verifying winter 2–5 hindcasts (as in Figures 5a, 6a, 7, 8, and 9). For some grid points the empirical first and second tercile are very close to each other or even identical (especially in regions of very low climatological winter storm frequency). This leads to ambiguous class definitions. Consequently, we excluded all grid points (additionally to those already neglected because of inconsistencies between ERA40 and ERA-Int.) exhibiting an inter-tercile range of less than 1 (for all levels of temporal aggregation) from skill assessments (also masked in gray for RPSS-figures; analogous to Kruschke et al., 2014).

For a given winter (or multi-winter mean) the reanalysis provides one certain category as observational reference (a probability of 1 for this category). The fraction of model ensemble members falling into a category yields the respective forecast probability. We consider cumulative probabilities for the three classes and hence use the *Ranked Probability Score* (RPS) as probabilistic verification measure. We apply an estimator of the RPS, developed by Ferro (2007, see also Ferro et al., 2008) and adapted by Kruschke et al. (2014) to account for ensemble size varying over initializations (as in the case of *baseline0*) in order to eliminate the systematic ensemble-size-dependent bias of the RPS:

$$\mathrm{RPS}_{\tau,M} = \frac{1}{I}\sum_{i=1}^{I}\sum_{k=1}^{K}(F_{\tau,i,k} - O_{t,k})^2 - \frac{M - m_i}{M(m_i - 1)}F_{\tau,i,k}(1 - F_{\tau,i,k}) \tag{2.1}$$

$F_{\tau,i,k}$ is the cumulative forecast probability of class k (with $K = 3$) derived from the forecast ensemble of initialization i (with $I = 41$) for a specific forecast lead time τ. $O_{t(i,\tau),k}$ is the cumulative probability of class k from observations for the time $t(i, \tau)$, corresponding to the time of initialization and forecast lead time. $O_{t(i,\tau),k}$ effectively is the Heaviside step function with $O_{t(i,\tau),k} = 1$ if class k or lower is observed and $O_{t(i,\tau),k} = 0$ otherwise. The second term of the equation constitutes the bias-correction, subtracting the systematic RPS-bias, an ensemble of size m_i suffers from, when compared to an ensemble of size M.

The benefit of a forecast compared to a reference forecast is quantified by the *Ranked Probability Skill Score* (RPSS)

$$\mathrm{RPSS}_{\tau} = 1 - \frac{\mathrm{RPS}_{\mathrm{fc},\tau}}{\mathrm{RPS}_{\mathrm{ref},\tau}}. \tag{2.2}$$

The performance of the initialized decadal hindcasts is assessed by using two different reference forecasts. The very basic reference is climatology. In the context of probabilistic prediction of three discrete classes this means climatological cumulative category probabilities ($\frac{1}{3}$, $\frac{2}{3}$, and 1) which are used as reference forecasts for all hindcasts and lead times. When assessing the skill compared to uninitialized simulations, it is again the fraction of (uninitialized) model ensemble members which is transfered to reference forecast probabilities for the three categories and the given time.

Significance (p-value < 5 %) of calculated skill scores is assessed by 1000-fold overlapping-block-bootstrapping (Künsch, 1989) from the available 41 hindcast-observation pairs.

3 Systematic model deviations: climatology and long-term trend

As already mentioned in Section 1, systematic deviations of the model simulations from observational reference might pose serious challenges for verification. The most problematic issues in this respect are model biases that are not constant over time. These may result from modeled externally forced long-term trends differing from observations (possible issue for uninitialized simulations as well as initialized predictions) or from

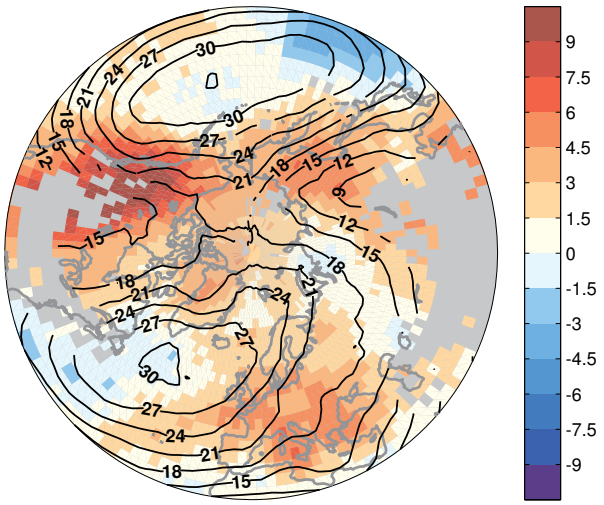

Figure 1: Climatological winter storm frequency (number of tracks per ONDJFM and 1000 km radius) from regression analysis (offset) for ERA-reanalyses (black contours) and biases of uninitialized simulations of MPI-ESM-LR (colored), calculated over winters 1961/62–2009/10; areas of strong inconsistencies between ERA40 and ERA-Int. are masked out (gray)

initialization states not compatible with the model's climatology. The latter will lead to (not necessarily monotonic) drifts of the model towards its climatology. It is important to estimate these different components of a model bias and properly separate them.

This section is dedicated to the assessment of the model's climatology and long-term trend in comparison to ECMWF's reanalyses. We assume that both, climatology and trend, are comparable for the uninitialized simulations and the initialized hindcasts. As already mentioned, the latter may additionally suffer from drifts but these are to be addressed in Section 4. Therefore, the regression analysis with a linear trend in time ($N = N_0 + N_t \cdot t$) for the annual winter storm frequencies (N) presented in this section are based on the uninitialized simulations of MPI-ESM-LR and the ERA-reanalyses only. Note that trend estimates are generally a result from external forcing and potential aliasing effects from multi-decadal natural variability. Small sample sizes – as for the reanalyses and the comparably short observational record – may also lead to artificial non-zero trends. For the ensemble of ten uninitialized simulations the sample size is sufficiently large and the influence of internal multi-decadal variability on the trend will be averaged out as these ten transient simulations represent different phases of natural low-frequency variabilities.

The offset parameters (N_0) derived for the linear trend (Figure 1) are equivalent to climatological winter storm frequencies over the analyzed period (winters 1961/62–2009/10), while the slope parameters (N_t, Figure 2) naturally estimate a linear change in this respect. Climatological NH winter storm frequencies are highest over the oceans and considerably lower over the continents (black contours in Figure 1 denote absolute numbers from ERA-reanalyses). Winter storm frequencies over the North Pacific are slightly higher than over the North Atlantic and the overall patterns are in very good agreement with other studies on extra-tropical cyclones and storm track diagnostics (see e.g. Ulbrich et al., 2008; Ulbrich et al., 2009). The differences between model and reanalyses (color shadings) reveal patterns very similar to those reported by Kruschke et al. (2014, see their Figure 1b) for cyclones instead of wind storms. The North Atlantic storm track is too zonal in the model with a slightly negative bias in the core of the storm track and positive deviations along the southern edge, especially over Europe. MPI-ESM-LR overestimates winter storm genesis over the Mediterranean. The model's North Pacific storm track is shifted northward (especially over the Northwest Pacific) and extends too far over North America, resulting into local winter storm frequencies up to more than 50 % higher than those from reanalyses.

The slope estimates (Figure 2) reveal several interesting features regarding linear changes of NH winter storm frequencies over the considered period. According to ERA-reanalyses (Figure 2(a)), the North Pacific, as well as the mid-latitudinal North Atlantic and the polar latitudes are dominated by significantly (95 %-confidence intervals of slope estimates not including zero) positive trends. Trends over the North Pacific seem more pronounced than those over the North Atlantic. The strongest trends in reanalysis data can be found in the Arctic. On the other hand, the Eurasian continent exhibits predominantly negative trends.

The uninitialized simulations of MPI-ESM-LR show considerably weaker trends for most regions (Figure 2(b)). At least in part, this is likely to be a result of sampling effects by using the full model ensemble compared to the single "realization" provided by the reanalyses. For this reason we focus only on significant trends and the accordance of their sign to reanalyses. In this respect model and reanalysis agree for the significantly positive trend over the Arctic. The significantly negative trend over Eastern Europe and Russia is also picked up by the model, even its magnitude is confirmed. However, this trend pattern is slightly shifted to the North. Additionally, MPI-ESM-LR features significantly negative trends over North America, which cannot be found in the reanalyses. For large parts of the North Atlantic and North Pacific – differing from reanalyses – MPI-ESM-LR exhibits no significant trends.

Both features tackled here – a climatological bias as well as deviating long-term trends of the model – are of relevance for decadal predictions and their verification. Climatological bias patterns as depicted in Figure 1 prove systematic misrepresentations like shifts or deformations of relevant variability patterns and phenomena. These pose serious issues to the interpretation and verification of predictions regarding the respective phenomenon and its implications. Even if the model was able to perfectly predict the temporal evolution of a specific feature, e.g. the North Atlantic storm track activ-

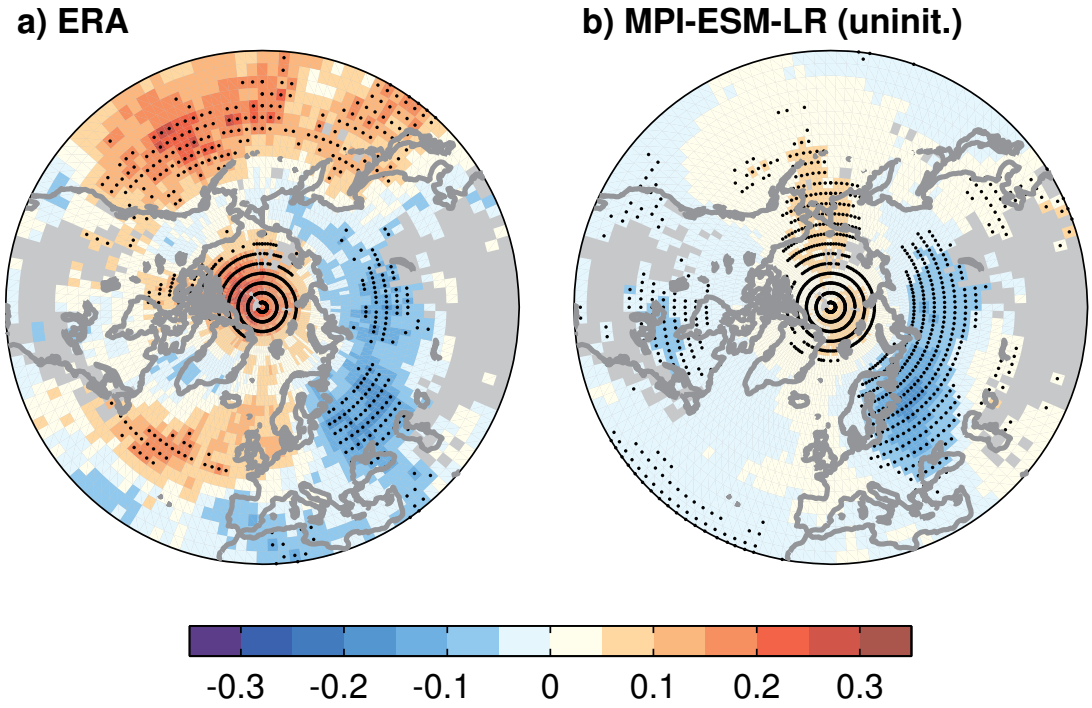

Figure 2: Linear change of winter storm frequency per year from regression analysis (slope) calculated over winters 1961/62–2009/10 for a) ERA-reanalyses and b) uninitialized simulations of MPI-ESM-LR; Trend significance (95 %-confidence intervals of slope estimates not including zero) marked as black dots; areas of strong inconsistencies between ERA40 and ERA-Int. are masked out (gray)

ity, its deterioration results into related signals at locations differing from observations. This hampers straightforward use of the predictions and their verification in the sense of grid-point-wise comparisons.

The reasons for trends, differing between reanalyses and model ensemble are threefold. First, part of the trends obtained for reanalysis data are probably due to internal multi-decadal variability with associated periodicities and phases such that they mimic a trend over the available observational period. As already stated, this part of temporal variability is assumed to be averaged out using the full ensemble of uninitialized simulations. The initialized hindcasts, on the other hand, may be able to (at least partially) pick up this component. Second, externally driven variability will be evident for both, reanalyses and model ensemble but may be imperfectly represented by the latter. Third, the very limited sample size for reanalysis data hampers a robust quantitative estimate of trends. We suppose that most of the major trend discrepancies between model (Figure 2(b)) and reanalyses (Figure 2(a)), namely those over the ocean basins, are a result of such multi-decadal internal variabilities appearing as linear trends for reanalyses over the analyzed period. This is based on the assumption that the model generally shows an appropriate response to external forcing (at least the same sign) with respect to the frequency of winter storms. Nevertheless, any imperfect representation (over- or underestimation) of externally driven long-term trends poses a serious constraint to the model's predictive skill: it misses one component relevant for climate evolution on decadal time scales. Furthermore, if the long term trend differs between model and reanalysis, the model drift after initialization is a function of time. This needs to be accounted for and will be addressed in Section 4.

4 Statistical adjustment of model bias, long-term trend, and hindcast drifts

According to the recommendation of the International CLIVAR Project Office (ICPO, 2011) a climatological bias is to be subtracted from anomaly-initialized predictions and uninitialized simulations, while for full-field-initialized predictions, the lead-time dependent bias is subtracted to account for probable model drifts over forecast lead time (implicitely also adjusting a climatological bias). Kharin et al. (2012) showed that this approach of assuming a model drift being constant for all initialization times is problematic, especially in the presence of long-term climate change signals differing between model and observations. We illustrate (Figure 3) this issue with respect to winter storm frequencies using the example of the *ORAff*-hindcasts (all 10 ensemble members) and a lead-time dependent bias calculated from the first 20 (1961–1980) and the last 20 (1982–2001) initializations, respectively. The recommendation of ICPO (2011) would be justified if both results exhibit no significant differences. However, model bias and its evolution with lead time (results for winter 1, 5 and 9 shown in Figure 3) substantially differ

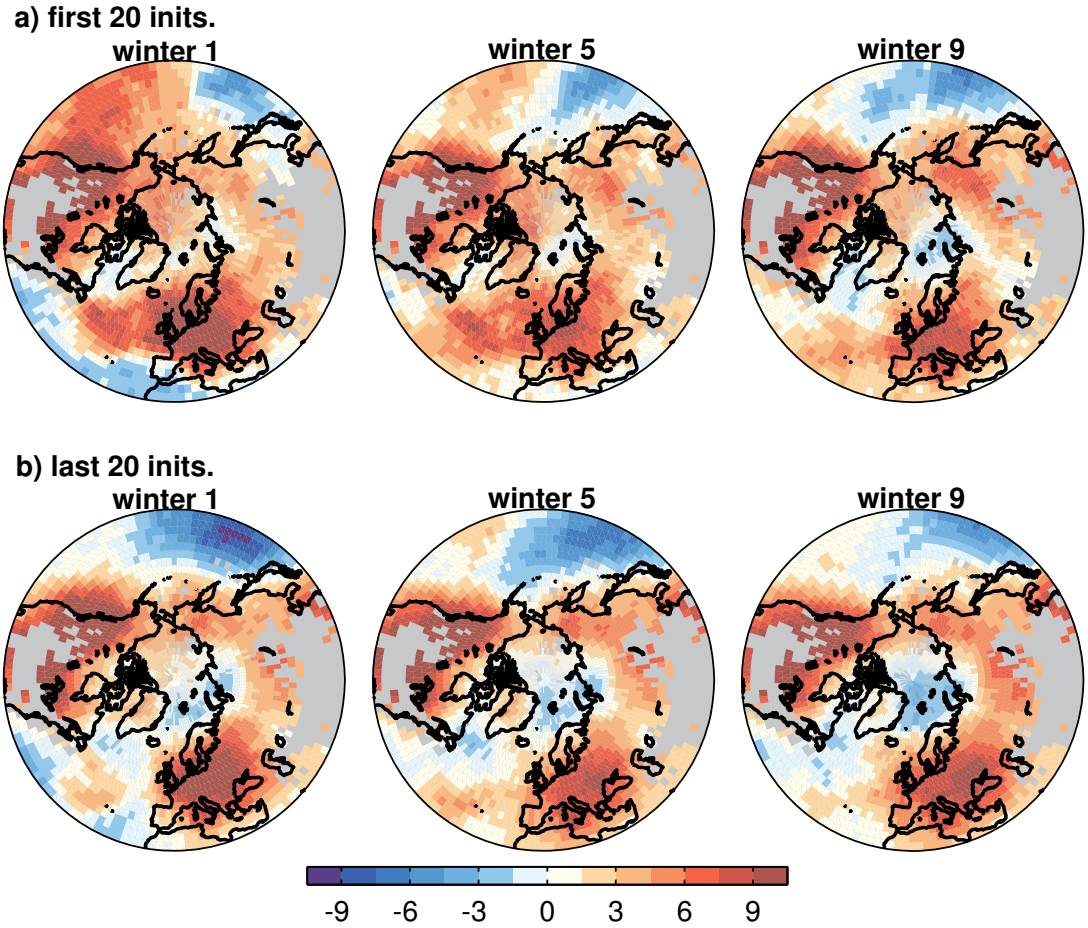

Figure 3: Lead-time-dependent bias of *ORAff*-hindcasts with respect to NH winter storm frequency (number of winter storm tracks per ONDJFM within 1000 km radius) with ERA-reanalyses as reference: a) calculated from first 20 initializations (1961–1980) only; b) calculated from last 20 initializations (1982–2001) only; areas of strong inconsistencies between ERA40 and ERA-Int. are masked out (gray)

between the earlier (upper panel in Figure 3) and later (lower panel in Figure 3) initialization times, especially for shorter lead times. While the model bias over the Northeast (Northwest) Pacific is positive (negative) during the first hindcast winters of the first 20 initializations and evolves towards neutral (more negative) conditions over lead time, it starts neutral (more negative) and shows no clear (positive) trends over lead time if calculated from the last 20 initializations. Over the North Atlantic, a strongly positive (slightly negative) bias is obvious over the mid-latitudes (Subtropics) during the first winters but evolving negatively (positively) over lead time, when calculated from the first 20 initializations. Based on the last 20 initializations, the temporal evolution of the bias over the subtropical North Atlantic is generally similar to that analyzed from the first 20 initializations, while no remarkable drifts are found over the mid-latitudinal North Atlantic with a bias constant over lead time and slightly negative here. A two-sided t-test shows significant (p-value < 0.05) differences between early and later initializations regarding winter 1 for most regions of the NH, including both stormtracks (not shown here). The corresponding differences for winter 9 are less pronounced, fewer regions exhibit significant disagreements. In fact, both winter 9 bias patterns are generally similar to the climatological bias pattern of the uninitialized simulations (see Figure 1), which confirms the general expectation of the initialized hindcasts to drift towards model climatology over lead-time.

We conclude, that we have to account for model drifts that are changing over time. Taking up the suggestion of Gangstø et al. (2013), we choose a parametric approach to account for the drift along lead time τ using a third order polynomial

$$\widehat{H}_{i,\tau,j} = H_{i,\tau,j} - a_0 - a_1\tau - a_2\tau^2 - a_3\tau^3. \qquad (4.1)$$

To account for the non-stationarity of the model drifts, we allow the polynomial parameters $a_k, k = 0, \ldots, 3$ to change over time t, i.e. $a_k = a_k(t)$. The most simple model is a linear trend in time as suggested by Kharin et al. (2012), i.e. $a_0 = b_0 + b_1 t$. This means, that for a certain initialization i and lead time τ the corrected

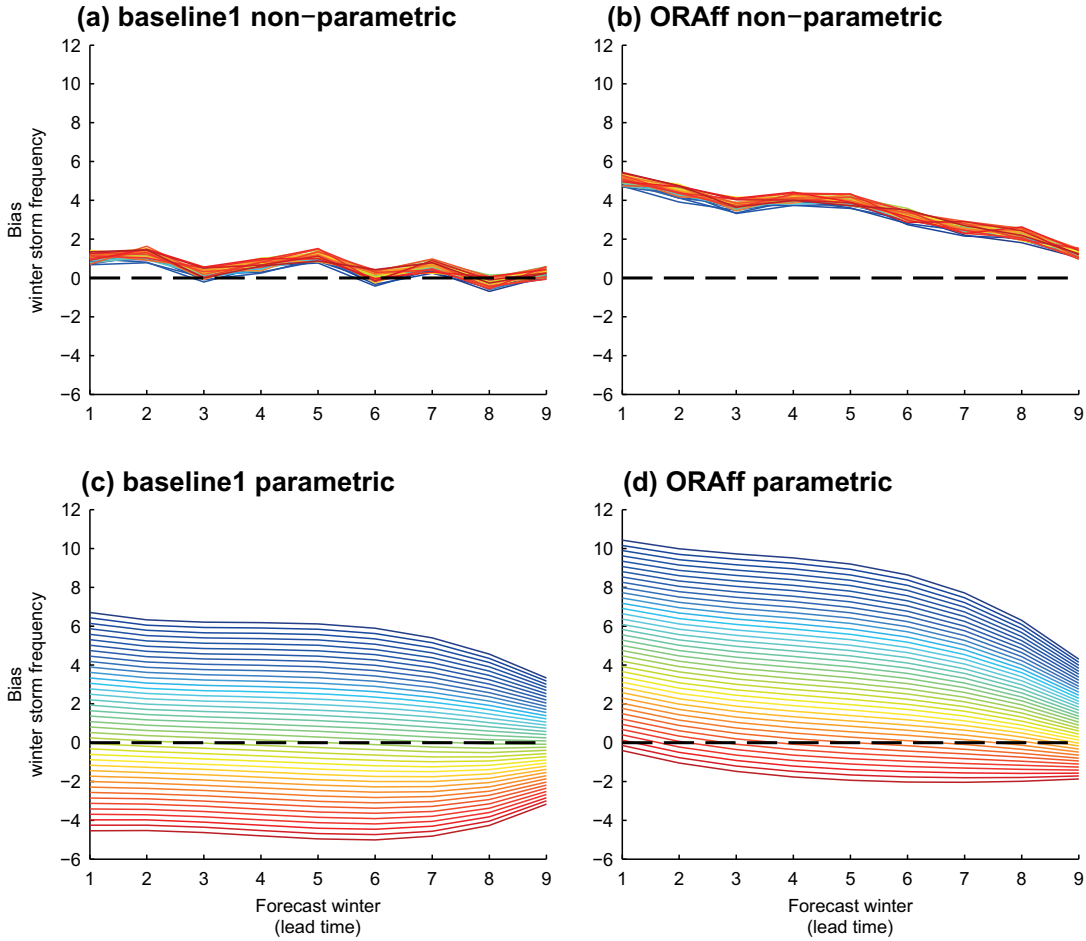

Figure 4: Lead-time-dependent bias (ERA-reanalyses as reference) of the 41 *baseline1*- (left) and *ORAff*-hindcasts (right) regarding winter storm frequency over the central North Atlantic (−30 ° E, 48.75 ° N), estimated with standard non-parametric approach in cross-validated manner (leaving out the respective hindcast, the bias is calculated for; top) and parametric approach applied in this study (bottom); line color denoting initialization time with blue for early hindcasts (starting 1961) and red for most recent hindcasts (ending 2001)

hindcast of the j^{th} ensemble member is given by

$$\widehat{H}_{i,\tau,j} = H_{i,\tau,j} - (b_0 + b_1 t) - (b_2 + b_3 t)\tau - (b_4 + b_5 t)\tau^2 - (b_6 + b_7 t)\tau^3 \quad (4.2)$$

with $H_{i,\tau,j}$ being its uncorrected equivalent.

The parameters $b_0 .. b_7$ are estimated by the standard least-squares-method from the differences between all available hindcasts (the individual ensemble members) and the reanalysis valid for the respective time t, corresponding to the given initialization and lead time. Compared to the previously described estimation of a separate bias for all of the nine hindcast winters, we have here only eight parameters to estimate instead of nine. The model is thus more parsimonious, nevertheless accounting for non-stationary drifts.

Figure 4 shows the effect of this parametric drift assessment in comparison to the non-parametric standard procedure (ICPO, (2011)) for the *baseline1*- and *ORAff*-hindcasts and the winter storm frequency, calculated for an exemplary grid point over the central North Atlantic (−30 ° E, 48.75 ° N). Please note, that the bias shown here is not the final product of different adjustment procedures. These are different estimates of what needs to be adjusted. Already visible from the non-parametric estimation of a lead-time-dependent bias (Figure 4(a) and (b)), a drift over lead-time is more obvious for *ORAff* than for the anomaly-initialized *baseline1* (for this grid point). This matches the general expectation of more pronounced drifts of full-field-initialized systems. The bias estimations resulting from the parametric approach (Figure 4(c) and (d)) now clarify how bias and drifts are changing over time. Closely related to the trend differences already seen in Figure 2, bias estimations for winter 1 and *baseline1* (*ORAff*) range from above 6 (10) for the earliest initialization to below −4 (0) for the latest initialization. Again the *baseline1*-bias shows comparably little change over lead-time, that is no substantial drifts are evident, while the *ORAff*-bias decreases remarkably, though not completely reaching the level of *baseline1*. Interestingly, the bias of the uninitialized simulations for the respective time (not shown here explicitly) is smaller than the *ORAff*-bias, too. This result suggests (not tested for statistical significance) that for winter storm frequencies over the North Atlantic, a simulation of approximately nine years is not long enough

to get completely rid of the full-field initialization shock. Also note that the range of the bias over different initialization times (Figure 4(c,d), different colors) is larger than its change over lead time. We conducted similar analyses comparing the raw differences (i.e. hindcast errors) for several grid points with those estimated by our parametric approach in order to qualitatively assess potential overfitting. No obvious indication of such an issue could be found for any of the grid points.

A correction of the hindcasts $H_{i,\tau,j}$ with this parametric approach effectively i) eliminates a climatological bias of the model (b_0 in Eq. (4.2)), ii) corrects for a deviating long-term trend linear in time ($b_1 t$ in Eq. (4.2)), and iii) removes a potential cubic drift with parameters varying linearly in time (terms including τ in Eq. (4.2)). Assessing the skill of the initialized hindcasts adjusted this way by comparing them to the raw uninitialized simulations would be unfair. Consequently, we have to adjust the latter as well. A straightforward application of Eq. (4.2) for the uninitialized simulations is not possible as the forecast lead time τ is not defined in their case. That leaves us with the adjusted "forecast" of an uninitialized simulation $\widehat{U}_{t,j}$ which is derived from

$$\widehat{U}_{t,j} = U_{t,j} - (c_0 + c_1 t) \tag{4.3}$$

based on the unadjusted "forecast" $U_{t,j}$. The climatological bias parameter c_0 is exactly that depicted as color shadings in Figure 1 and the linear trend parameter c_1 for the uninitialized simulations is the difference between Figure 2(a) and (b).

We individually apply these corrections to the winter storm frequencies calculated for all grid points (Eq. (4.2) to the initialized hindcasts, Eq. (4.3) to the uninitialized simulations). Kharin et al. advised against using this approach for localized quantities as local trends derived from observations exhibit large uncertainties. They suggested a simple approach (for surface temperature) by correcting local model trends with the ratio of global trends from model and observations, which are more reliable for the latter than their local peculiarity. The current understanding of long-term climate change signals with respect to storm tracks and typical pathways of intense extra-tropical cyclones and winter storms (summarized in Hartmann et al., 2013; Christensen et al., 2013) is more characterized by shifts and local trends than by globally uniform changes. Hence, such an approach is not appropriate in the context of our study. Here, the winter storm frequencies are derived for a high degree of spatial aggregation (counting numbers of tracks within a radius of 1000 km around the given grid-point) and thus these values are very different from, e.g., grid-box-wise temperatures. We thus assume that the respective values derived from reanalyses are robust enough to be directly used for estimating the parameters of the time dependent cubic drift correction. This is supported by trend patterns calculated for the *Twentieth Century Reanalysis* (20CR, Compo et al., 2011) being in very good agreement with those of the ERA-reanalyses, only slightly weaker with respect to the maxima (not shown).

5 Decadal prediction skill

The multiple-system ensembles deliver more robust estimates of skill because of their larger ensemble sizes (compared to the individual hindcast experiments). Thus, we address our first two research questions regarding skill over climatological forecasts and the benefit from initialization by means of the *Grand Ensemble*.

To answer the general question, whether decadal predictions contain any valuable information about the winter storm frequency of the upcoming years, we compare the results of the *Grand Ensemble* to climatological forecasts. The forecasts of average winter storm frequency for winters 2–5 (Figure 5(a)) exhibit significantly positive skill for the entire Pacific basin, the NH polar latitudes, as well as the mid-latitudinal North Atlantic and a region over the Mediterranean and the Black Sea. For winters 2–9 (Figure 5(b)) skill scores are even higher for most parts of the NH. The only region contiguously exhibiting zero or slightly negative skill scores is the Atlantic sector of the Arctic ocean and adjacent land areas, such as Greenland, Scandinavia and parts of Russia. Over the North Pacific, skill seems to be concentrated in the mid-latitudes. Skill over the central sub-tropical North Pacific is slightly lower for winters 2–9 than for the winter 2–5 forecasts.

The general result from these analyses is very encouraging. However, two major issues have to be stated: First, the initialized decadal predictions (and the uninitialized runs) were adjusted to match the linear trend of the observations. So the detected skill might be a result of this trend adjustment only. Second, the results shown so far are not sufficient to prove the success of initialized decadal predictions. At least part of the skill compared to a climatological forecast could arise from the response to external forcing, already included in the uninitialized simulations.

To falsify the first issue, confirming that these positive findings are not purely a result from the statistical adjustment, we calculated the skill of the raw (not trend-adjusted) uninitialized simulations (4 yr- and 8 yr-running-means being the equivalents to the 2–5 yr- and 2–9 yr-forecasts) over climatological forecasts. We found that these raw uninitialized runs do exhibit skill patterns (not shown) generally similar to those of the initialized hindcasts as in Figure 5. While the magnitudes of these skill scores for the 4 yr-running-mean is overall considerably lower than those of the initialized 2–5 yr-forecasts (Figure 5(a)), RPSS-magnitudes for the 8 yr-running-mean are comparable to those of the initialized 2–9 yr-forecasts (Figure 5(b)). Thus, the statistical adjustment is not the major source of skill in this context but this skill of the uninitialized simulations emphasizes the second issue regarding external forcing as the source of skill.

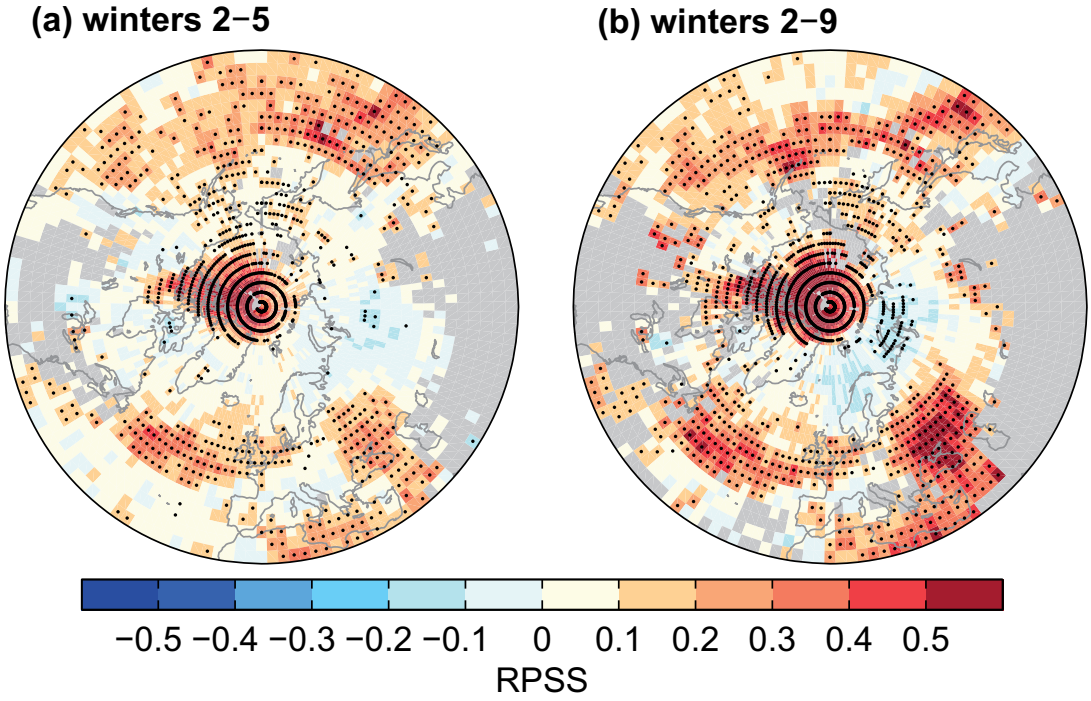

Figure 5: RPSS of *Grand Ensemble* over climatological forecasts regarding the average winter storm frequency (number of tracks per ONDJFM in the vicinity of 1000 km) for (a) hindcast winters 2–5 and (b) hindcast winters 2–9 based on ERA-reanalyses as observational reference; significant skill scores ($\alpha < 5\,\%$) as black dots, areas of either strong inconsistencies between ERA40 and ERA-Int. or ambiguous event class definitions are masked out (gray)

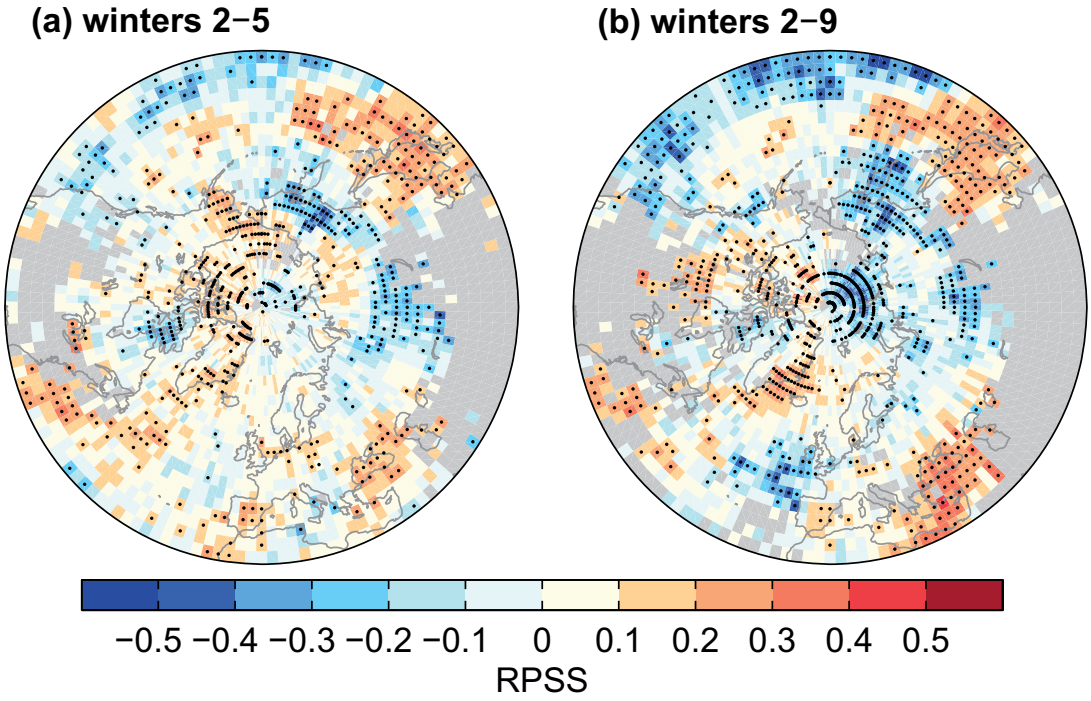

Figure 6: RPSS of *Grand Ensemble* over uninitialized simulations regarding the average winter storm frequency (number of tracks per ONDJFM in the vicinity of 1000 km) for (a) hindcast winters 2–5 and (b) hindcast winters 2–9 based on ERA-reanalyses as observational reference; significant skill scores ($\alpha < 5\,\%$) as black dots, areas of strong inconsistencies between ERA40 and ERA-Int. are masked out (grey)

To answer the question about the added value of initialization, skill scores are additionally calculated with the (trend-adjusted) uninitialized runs as reference forecast. Figure 6(a) shows the results for the winter 2–5 forecast of the *Grand Ensemble*. Obviously, only for some regions the initialized hindcasts are able to provide significantly added value. The most prominent example is the entrance of the North Pacific storm track over Eastern Asia and the Northwest Pacific. Similarly but less pronounced and coherent, decadal predictions for winter storm frequencies at the entrance of the North Atlantic storm track along the North-American east coast seem to profit from initialization. The only other region where significantly positive skill scores can be diagnosed over a larger area is the American sector of the Arctic Ocean. These results are generally in line with those for intense cyclones investigated in the study of Kruschke et al. (2014). In most cases we find areas of significantly positive skill for winter wind storm frequency south of areas exhibiting skill with respect to intense cyclone frequency (Kruschke et al., 2014, Figure 3c and 5c). This matches our expectation, as the winter storm tracks, diagnosed by tracking wind extremes, are usually found south of the related extra-tropical cyclone track. Skill of winter 2–5 predictions over the North Pacific storm track is, however, smaller than values found by Kruschke et al.. This may be – at least partially – explained by the different observational references used, as Kruschke et al., found a strong influence on the skill from the specific reanalysis dataset used. They state that analogous analyses regarding cyclone frequencies based on NCEP1-reanalysis (Kalnay et al., 1996) or a mix of ERA40 and ERA-Interim, as done in this paper, wipe out the skill they found in this area, using 20CR as observational reference. Considering predictions of the average winter storm frequency of winters 2–9 (Figure 6(b)), only the positive skill pattern of the winter 2–5 forecast over the North-West Pacific prevails. On the other hand, an area over the Eastern Mediterranean and the Black Sea is marked by significantly positive skill scores for these forecast horizons. For all other regions of the NH, that is the subtropical North Pacific, the North-East Atlantic, large parts of Eurasia and the adjacent Arctic Ocean, our initialized decadal predictions are not able to provide skill over the uninitialized transient simulations.

To systematically evaluate the different initialization strategies followed so far, we compare the multiple system ensembles *ano-Ens.* and *ff-Ens.* as well as *ORA-Ens.* and *GECCO-Ens.* (only skill over uninitialized simulations for winter 2–5 hindcasts shown in Figure 7 and 8, respectively). Generally, the differences with respect to the calculated skill scores between initializing from oceanic full-fields or anomalies are rather small. *ff-Ens.* seems to perform slightly better than *ano-Ens.*, though not significant, in predicting winter storm frequencies over the Northwest Pacific, while anomaly-initialization yields better results over the Arctic Ocean north of America and around the Black Sea. Most remarkable about the comparison of *ORA-Ens.* and *GECCO-Ens.* are the better predictions of the latter regarding winter storm frequencies over the central and western North Pacific. The same seems to be the case over the Northwest Atlantic. Both, *ORA-Ens.* and *GECCO-Ens.* perform well over the American Arctic, but the area of significant skill is more coherent for *ORA-Ens.*. Overall, none of the initialization strategies is clearly superior to the others.

Inspired by the studies of Kharin et al. (2012) and Gangstø et al. (2013) we introduced a novel approach to assess potential model drifts over lead time in Section 4. Figure 9 is dedicated to evaluating the effect of this approach in comparison to the standard non-parametric procedure by way of example for the winter 2–5 forecasts of the *ORAff* hindcast experiment. The standard approach recommended by ICPO (2011, Figure 9(a)) implicitly corrects only for a climatological bias while our parametric approach (Figure 9(b)) additionally adjusts the linear long-term trend. In order to evaluate the impact of drift assessment only we adjusted the uninitialized simulations used as reference forecasts accordingly for this comparison. That means bias-correction only for Figure 9(a); bias- and trend-correction for Figure 9(b). Subsequently, differences between Figure 9(a) and (b) are due to the more or less appropriate approach regarding drift assessment only. These differences clearly show a generally higher skill of *ORAff*, when adjusted with the parametric drift-correction approach. The overall pattern is not changed and local skill maxima/minima stay the same, though. The superiority of our parametric approach is evident for all individual hindcast experiments as well as the multiple-system ensembles (not shown). However, the effect is more pronounced in the presence of substantial drifts over lead time, that is for full-field initialized ensembles.

6 Summary and discussion

Five sets of decadal hindcasts produced within the MiKlip initiative are analyzed with respect to the skill of probabilistic three-category predictions regarding winter wind storm frequencies over the extra-tropical NH. Multiple-system ensembles are constructed by specific combinations of the original hindcast experiments to provide robust skill assessments (due to large ensemble sizes) and to permit systematic comparisons of different initialization strategies pursued so far.

It is shown that predictions of average winter storm frequency of winters 2–5 as well as winters 2–9 do exhibit significant skill (i.e. better than assuming climatological probabilities for each category and initialization; Figure 5) for large parts of the extra-tropical NH, that is the whole North Pacific, the mid-latitudinal North Atlantic, the American sector of the Arctic and a region over the Eastern Mediterranean and the Black Sea. This

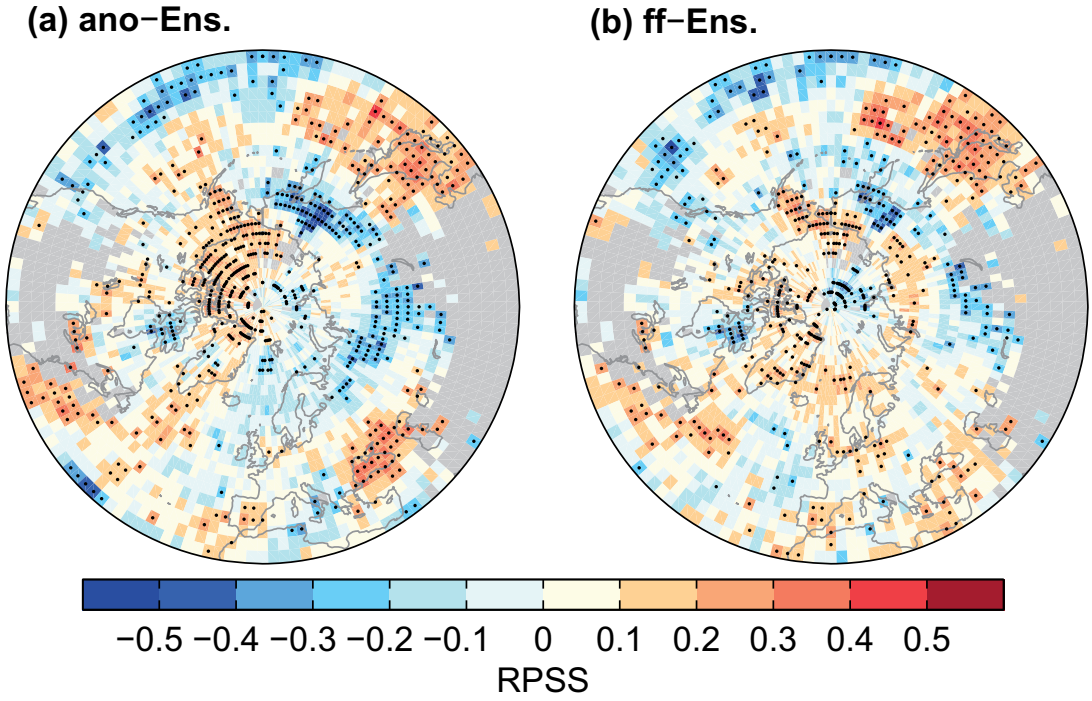

Figure 7: RPSS of (a) anomaly-initialized hindcasts (*ano-Ens.*) and (b) full-field-initialized hindcasts (*ff-Ens.*) over uninitialized simulations regarding the average winter storm frequency (number of tracks per ONDJFM in the vicinity of 1000 km) for hindcast winters 2–5 based on ERA-reanalyses as observational reference; significant skill scores ($\alpha < 5\,\%$) as black dots, areas of strong inconsistencies between ERA40 and ERA-Int. are masked out (grey)

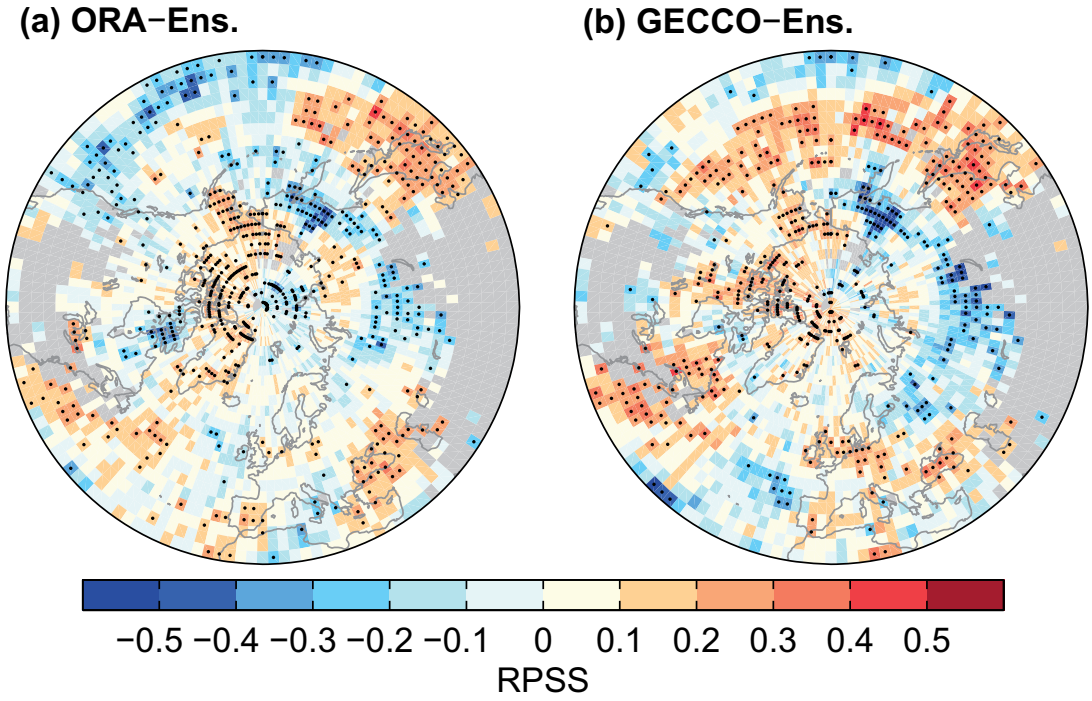

Figure 8: RPSS of hindcasts initialized from (a) ORA-S4- (*ORA-Ens.*) and (b) GECCO2-ocean-reanalysis (*GECCO-Ens.*) over uninitialized simulations regarding the average winter storm frequency (number of tracks per ONDJFM in the vicinity of 1000 km) for hindcast winters 2–5 based on ERA-reanalyses as observational reference; significant skill scores ($\alpha < 5\,\%$) as black/white dots, areas of strong inconsistencies between ERA40 and ERA-Int. are masked out (grey)

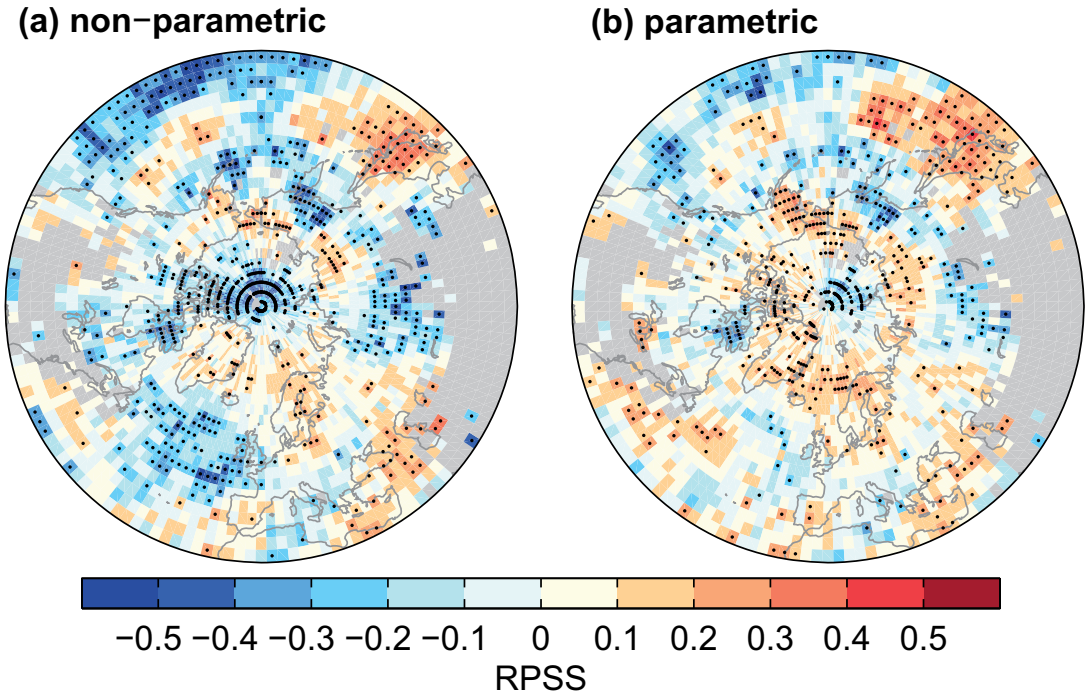

Figure 9: RPSS of *ORAff* over uninitialized simulations after (a) non-parametric and (b) parametric drift correction regarding the average winter storm frequency (number of tracks per ONDJFM in the vicinity of 1000 km) for hindcast winters 2–5 based on ERA-reanalyses as observational reference; significant skill scores ($\alpha < 5\,\%$) as black/white dots, areas of strong inconsistencies between ERA40 and ERA-Int. are masked out (grey)

result seems to be in line with the study of Haas et al. (accepted), analyzing prediction skill of the MiKlip system with respect to (right tail) quantiles of wind probability distributions over Europe.

However, a comparison to uninitialized (transient) simulations (Figure 6) reveals that a substantial part of this skill is attributed to long-term trends, associated with imposed greenhouse gas and aerosol forcing. Hence, the additional value of initialization effort is restricted to smaller areas. These are the entrance regions of both NH storm tracks along the North-American East Coast and especially over East Asia and the Kuroshio Extension. First analyses regarding the origin of this skill indicate a close relation to low-frequency variability of temperature gradients between (sub-)tropical water masses (trop. East Pacific to West Atlantic and the China Seas, respectively) and mid-latitude land areas (North America and Central to Northeastern Asia, respectively). The processes behind these patterns are subject of future research. Additionally, predictions for parts of the Arctic and the already-mentioned area over the Eastern Mediterranean and Black Sea profit from initialization. Regarding the latter region, skill for winters 2–9 is higher than for winters 2–5.

A systematic comparison of hindcasts produced by anomaly- vs. full-field-initialization (Figure 7) and initialized from ORA-S4- vs. GECCO2-ocean-reanalysis (Figure 8) yields no clearly superior initialization strategy. The only remarkable difference seems to be existent over the North Pacific, where *GECCO-Ens.* provides higher predictive skill than *ORA-Ens.* for both winter 2–5 and 2–9 forecasts (the latter not shown).

The skill of decadal predictions regarding winter wind storm frequency in the NH storm track regions is generally lower than skill regarding (intense) cyclone frequency as presented in the study of Kruschke et al. (2014). This is not caused by methodological differences. Applying this study's methods to correct for bias, long-term-trend and drift (Section 4) to the cyclone-related analyses of Kruschke et al. yields even higher skill of the initialized predictions over uninitialized simulations than presented in their study. As the vast majority of wind storms identified with the scheme applied here can be related to extra-tropical cyclones (not shown), it is not about completely different phenomena considered but about specific subsets of events. Kruschke et al. (2014) found promising results especially for intense cyclones' frequencies. Their definition of such events led to climatological frequencies in the storm track regions that are approx. twice as high as the climatological frequencies of winter storms (see Figure 1) in these regions as diagnosed in this paper. This means that the current study considers more extreme events in these regions than Kruschke et al. (2014). We assume that the inter-annual to decadal variability of this subset exhibits lower signal-to-noise-ratios (beyond externally forced trends) for large parts of the storm tracks than the less extreme subset of Kruschke et al. (2014), resulting in lower predictability and hence prediction skill.

We introduced a novel parametric approach to correct for potential model drifts, inspired by Kharin et al. (2012) and Gangstø et al. (2013). Comparison of skill assessments after using the standard drift-correction procedure recommended by ICPO (2011; also applied in Kruschke et al., 2014) and this parametric method (Figure 9), confirms the expectation that the latter is more adequate for estimating and subsequently eliminating model drifts. This leads to better skill assessments. Qualitative analyses gave no indication that this might be a result from overfitting by this method. However, this could be different for other variables exhibiting lower inter-annual variability and stronger serial correlation than winter storm frequencies which are analyzed in this study. A constraint of our parametric approach is the inherent adjustment of linear long-term trends. First, such linear long-term trends will not be suitable for each parameter, region and period. Second, it surely is disputable to which degree (or lead time) decadal predictions are reliable if the underlying model exhibits deviating long-term trends due to external forcing. Though, it is worth mentioning that our approach effectively is very similar to the widely used detrending in this respect. Third, a linear long-term trend is the most simple way to account for non-stationarity of climate and model bias; it is questionable if this approach is satisfactory in the more general context (e.g. for other variables) and with respect to the consequence of necessary extrapolations for future decadal forecasts. The general expectation is that the drift depends on the degree of "incompatibility" between initial conditions and the model's climatology, which is not necessarily a function of time as in our approach. However, a quantitative assessment of this incompatibility is not a trivial task. The recent study of Fuckar et al. (2014) is a first step in this direction but further research is needed to come up with optimal solutions regarding drift correction and the understanding of non-stationary model biases.

The earth system model in use, and particularly its atmospheric component is a critical element regarding forecast skill. ECHAM6 is as good as or even better than its predecessor in representing recent climate (Stevens et al., 2013). A particular finding is that the simulated extra-tropical circulation and stationary wave structure are closer to observations than in previous versions of the same model. Still, Stevens et al. diagnose an equatorward shift of the midlatitude jets. This particular bias is consistent with the model's zonalization of the North-Atlantic storm track and the northward shift of the North-Pacific storm track found in the current study (and by Kruschke et al., 2014). Such a bias can constitute a major issue for the assessment of prediction skill on a grid-point basis. Even with a perfect forecast of the temporal variability of these physically linked features, the location bias would lead to a mismatch with reality and thus to a deflation of the estimated prediction skill. As long as such systematic misrepresentations of storm track location and orientation are existent in model simulations, approaches accounting for this feature would be more appropriate.

Not neglecting these constraints, we nevertheless consider the results found as being encouraging. It is shown that initialized decadal predictions do provide potentially valuable information for several NH regions that go beyond externally forced long-term changes. Given the yet early stage of decadal prediction research and ongoing activities regarding model developments and improved initialization as well as adequate statistical post-processing and verification methods the prospects of furthermore improved decadal predictions regarding the frequency of winter storms and potentially related socio-economic value can be judged favorably.

Acknowledgments

We acknowledge funding from the Federal Ministry of Education and Research in Germany (BMBF) through the research program MiKlip (FKZ: 01LP1104A, 01LP1144A, 01LP1160A) and partly from Munich Re. We appreciate the constructive comments by Felicitas Hansen and two anonymous reviewers.

References

Balmaseda, M.A., K. Mogensen, A.T. Weaver, 2013: Evaluation of the ECMWF ocean reanalysis system ORAS4. – Quart. J. Roy. Meteor. Soc. **139**, 1132–1161, DOI: 10.1002/qj.2063.

Boer, G., V. Kharin, W. Merryfield, 2013: Decadal predictability and forecast skill. – Climate Dyn. **41**, 1817–1833, DOI: 10.1007/s00382-013-1705-0.

Christensen, J., K. Krishna Kumar, E. Aldrian, S.-I. An, I. Cavalcanti, M. de Castro, W. Dong, P. Goswami, A. Hall, J. Kanyanga, A. Kitoh, J. Kossin, N.-C. Lau, J. Renwick, D. Stephenson, S.-P. Xie, T. Zhou, 2013: Climate Phenomena and their Relevance for Future Regional Climate Change. – In: Stocker, T., D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, P. Midgley (Eds.): Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge Univ. Press, Cambridge, United Kingdom and New York, NY, USA, chapter 14, 1217–1308.

Compo, G.P., J.S. Whitaker, P.D. Sardeshmukh, N. Matsui, R.J. Allan, X. Yin, J. Gleason, B.E., R.S. Vose, G. Rutledge, P. Bessemoulin, S. Broennimann, M. Brunet, R.I. Crouthamel, A.N. Grant, P.Y. Groisman, P.D. Jones, M.C. Kruk, A.C. Kruger, G.J. Marshall, M. Maugeri, H.Y. Mok, O. Nordli, T.F. Ross, R.M. Trigo, X.L. Wang, S.D. Woodruff, S.J. Worley, 2011: The Twentieth Century Reanalysis Project. – Quart. J. Roy. Meteor. Soc. **137**, 1–28, DOI: 10.1002/qj.776.

Counillon, F., I. Bethke, N. Keenlyside, M. Bentsen, L. Bertino, F. Zheng, 2014: Seasonal-to-decadal predictions with the ensemble Kalman filter and the Norwegian Earth System Model: a twin experiment. – Tellus A **66**, 21074, DOI: 10.3402/tellusa.v66.21074.

Dee, D.P., S.M. Uppala, A.J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M.A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A.C.M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A.J. Geer, L. Haimberger, S.B. Healy, H. Hersbach, E.V. Holm, L. Isaksen, P. Kallberg, M. Koehler, M. Matricardi, A.P. McNally, B.M. Monge-Sanz, J.J. Morcrette, B.K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.N. Thepaut, F. Vitart, 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. – Quart. J. Roy. Meteor. Soc. **137**, 553–597, DOI: 10.1002/qj.828.

Doblas-Reyes, F.J., I. Andreu-Burillo, Y. Chikamoto, J. Garcia-Serrano, V. Guemas, M. Kimoto, T. Mochizuki, L.R.L. Rodrigues, G.J. van Oldenborgh, 2013: Initialized near-term regional climate change prediction. – Nat. Commun. **4**, 1715, DOI: 10.1038/ncomms2704.

Donat, M.G., D. Renggli, S. Wild, L.V. Alexander, G.C. Leckebusch, U. Ulbrich, 2011: Reanalysis suggests long-term upward trends in European storminess since 1871. – Geophys. Res. Lett. **38**, L14703, DOI: 10.1029/2011GL047995.

Eade, R., E. Hamilton, D.M. Smith, R.J. Graham, A.A. Scaife, 2012: Forecasting the number of extreme daily events out to a decade ahead. – J. Geophys. Res.-Atmos. **117**, D21110, DOI: 10.1029/2012JD018015.

Ferro, C.A.T., 2007: Comparing Probabilistic forecasting systems with the brier score. – Wea. Forecast. **22**(5), 1076–1088, DOI: 10.1175/WAF1034.1.

Ferro, C.A.T., D.S. Richardson, A.P. Weigel, 2008: On the effect of ensemble size on the discrete and continuous ranked probability scores. – Meteor. Appl. **15**, 19–24, DOI: 10.1002/met.45.

Fuckar, N.S., D. Volpi, V. Guemas, F.J. Doblas-Reyes, 2014: A posteriori adjustment of near-term climate predictions: Accounting for the drift dependence on the initial conditions. – Geophys. Res. Lett. **41**, 5200–5207, DOI: 10.1002/2014GL060815.

Gangstø, R., A.P. Weigel, M.A. Liniger, C. Appenzeller, 2013: Methodological aspects of the validation of decadal predictions. – Climate Res. **55**, 181–200, DOI: 10.3354/cr01135.

Garcia-Serrano, J., F.J. Doblas-Reyes, 2012: On the assessment of near-surface global temperature and North Atlantic multi-decadal variability in the ENSEMBLES decadal hindcast. – Climate Dyn. **39**, 2025–2040, DOI: 10.1007/s00382-012-1413-1.

Giorgetta, M.A., J. Jungclaus, C.H. Reick, S. Legutke, J. Bader, M. Böttinger, V. Brovkin, T. Crueger, M. Esch, K. Fieg, K. Glushak, V. Gayler, H. Haak, H.-D. Hollweg, T. Ilyina, S. Kinne, L. Kornblueh, D. Matei, T. Mauritsen, U. Mikolajewicz, W. Mueller, D. Notz, F. Pithan, T. Raddatz, S. Rast, R. Redler, E. Roeckner, H. Schmidt, R. Schnur, J. Segschneider, K.D. Six, M. Stockhause, C. Timmreck, J. Wegner, H. Widmann, K.-H. Wieners, M. Claussen, J. Marotzke, B. Stevens, 2013: Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the coupled model intercomparison project phase 5. – J. Adv. Model. Earth Syst. **5**, 572–597, DOI: 10.1002/jame.20038.

Goddard, L., A. Kumar, A. Solomon, D. Smith, G. Boer, P. Gonzalez, V. Kharin, W. Merryfield, C. Deser, S. Mason, B. Kirtman, R. Msadek, R. Sutton, E. Hawkins, T. Fricker, G. Hegerl, C. Ferro, D. Stephenson, G. Meehl, T. Stockdale, R. Burgman, A. Greene, Y. Kushnir, M. Newman, J. Carton, I. Fukumori, T. Delworth, 2013: A verification framework for interannual-to-decadal predictions experiments. – Climate Dyn. **40**, 245–272, DOI: 10.1007/s00382-012-1481-2.

Haas, R., M. Reyers, J.G. Pinto, 2015: Decadal predictability of regional-scale peak winds over Europe using the Earth System Model of the Max-Planck-Institute for Meteorology. – Meteorol. Z. **25**, 739–752, DOI: 10.1127/metz/2015/0583.

Hanlon, H.M., G.C. Hegerl, S.F.B. Tett, D.M. Smith, 2013: Can a Decadal Forecasting System Predict Temperature Extreme Indices?. – J. Climate **26**, 3728–3744, DOI: 10.1175/JCLI-D-12-00512.1.

Hartmann, D., A. Klein Tank, M. Rusticucci, L. Alexander, S. Brönnimann, Y. Charabi, F. Dentener, E. Dlugokencky, D. Easterling, A. Kaplan, B. Soden, P. Thorne, M. Wild, P. Zhai, 2013: Observations: Atmosphere and Surface. – In: Stocker, T., D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. Midgley (Eds.): Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. – Cambridge Univ. Press, Cambridge, United Kingdom and New York, NY, USA, chapter 2, 159–254.

Hawkins, E., B.W. Dong, J. Robson, R. Sutton, D. Smith, 2014: The Interpretation and Use of Biases in Decadal Climate Predictions. – J. Climate **27**, 2931–2947, DOI: 10.1175/JCLI-D-13-00473.1.

Hazeleger, W., V. Guemas, B. Wouters, S. Corti, I. Andreu-Burillo, F.J. Doblas-Reyes, K. Wyser, M. Caian, 2013: Multiyear climate predictions using two initialization strategies. – Geophys. Res. Lett. **40**, 1794–1798, DOI: 10.1002/grl.50355.

International CLIVAR Project Office, 2011: Data and Bias Correction for Decadal Climate Predictions. published online, compiled by CMIP-WGCM-WGSIP Decadal Climate Prediction Panel.

Jungclaus, J.H., N. Fischer, H. Haak, K. Lohmann, J. Marotzke, D. Matei, U. Mikolajewicz, D. Notz, J.S. von Storch, 2013: Characteristics of the ocean simulations in the Max Planck Institute Ocean Model (MPIOM) the ocean component of the MPI-Earth system model. – J. Adv. Model. Earth Syst. **5**, 422–446, DOI: 10.1002/jame.20023.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K.C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, D. Joseph, 1996: The NCEP/NCAR 40-year reanalysis project. – Bull. Amer. Meteor. Soc. **77**, 437–471, DOI: 10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Kharin, V.V., G.J. Boer, W.J. Merryfield, J.F. Scinocca, W.S. Lee, 2012: Statistical adjustment of decadal predictions in a changing climate. – Geophys. Res. Lett. **39**, L19705, DOI: 10.1029/2012GL052647.

Klawa, M., U. Ulbrich, 2003: A model for the estimation of storm losses and the identification of severe winter storms in Germany. – Nat. Hazard Earth Sys. **3**, 725–732.

Köhl, A., 2015: Evaluation of the GECCO2 ocean synthesis: transports of volume, heat and freshwater in the Atlantic. – Quart. J. Roy. Meteor. Soc. **141**, 166–181, DOI: 10.1002/qj.2347.

Kröger, J., W. Müller, J.-S. von Storch, 2012: Impact of different ocean reanalyses on decadal climate prediction. – Climate Dyn. **39**, 795–810, DOI: 10.1007/s00382-012-1310-7.

Kruschke, T., 2014: Winter wind storms: Identification, verification of decadal predictions, and regionalization. – Ph.D. thesis, Institute of Meteorology – Dept. of Earth Sciences – Freie Universität Berlin, Berlin, Germany.

Kruschke, T., H.W. Rust, C. Kadow, G.C. Leckebusch, U. Ulbrich, 2014: Evaluating decadal predictions of northern hemispheric cyclone frequencies. – Tellus A **66**, 22830, DOI: 10.3402/tellusa.v66.22830.

Künsch, H.R., 1989: The Jackknife and the Bootstrap For General Stationary Observations. – Annals of Statistics **17**, 1217–1241, DOI: 10.1214/aos/1176347265.

Leckebusch, G.C., U. Ulbrich, L. Froehlich, J.G. Pinto, 2007: Property loss potentials for European midlatitude storms in a changing climate. – Geophys. Res. Lett. **34**, L05703, DOI: 10.1029/2006GL027663.

Leckebusch, G.C., D. Renggli, U. Ulbrich, 2008: Development and Application of an Objective Storm Severity Measure for the Northeast Atlantic Region. – Meteor. Z. **17**, 575–587.

Matei, D., H. Pohlmann, J. Jungclaus, W. Müller, H. Haak, J. Marotzke, 2012: Two Tales of Initializing Decadal Climate Prediction Experiments with the ECHAM5/MPI-OM Model. – J. Climate **25**, 8502–8523, DOI: 10.1175/JCLI-D-11-00633.1.

Meehl, G.A., H. Teng, 2014: CMIP5 multi- model hindcasts for the mid-1970s shift and early 2000s hiatus and predictions for 2016–2035. – Geophys. Res. Lett. **41**, 1711–1716, DOI: 10.1002/2014GL059256.

Meehl, G.A., L. Goddard, J. Murphy, R.J. Stouffer, G. Boer, G. Danabasoglu, K. Dixon, M.A. Giorgetta, A.M. Greene, E. Hawkins, G. Hegerl, D. Karoly, N. Keenlyside, M. Kimoto, B. Kirtman, A. Navarra, R. Pulwarty, D. Smith, D. Stammer, T. Stockdale, 2009: Decadal Prediction – Can It Be Skillful?. – Bull. Amer. Meteor. Soc. **90**, 1467–1485, DOI: 10.1175/2009BAMS2778.1.

Meehl, G.A., L. Goddard, G. Boer, R. Burgman, G. Branstator, C. Cassou, S. Corti, G. Danabasoglu, F. Doblas-Reyes, E. Hawkins, A. Karspeck, M. Kimoto, A. Kumar, D. Matei, J. Mignot, R. Msadek, H. Pohlmann, M. Rienecker, T. Rosati, E. Schneider, D. Smith, R. Sutton, H. Teng, G.J. van Oldenborgh, G. Vecchi, S. Yeager, 2014: Decadal Climate Prediction: An Update from the Trenches. – Bull. Amer. Meteor. Soc. **95**, 243–267, DOI: 10.1175/BAMS-D-12-00241.1.

Munich RE Group, 2008: Knowledge series: Highs and lows – Weather risks in central Europe. – Published online, e.g. http://www.mroc.com/publications.html.

Munich RE Group, 2013: Topics Risk Solutions – Snow, freezing rain and Arctic temperatures. – Published online http://www.munichre.com/site/wrap/get/documents_E-506421131/mram/assetpool.munichreamerica.wrap/PDF/03Weather\T1\textbackslash%20Risks/TRS_4_2013_nathazards_snow.pdf, Princeton, NJ, USA.

Müller, W.A., J. Baehr, H. Haak, J.H. Jungclaus, J. Kröger, D. Matei, D. Notz, H. Pohlmann, J.S. von Storch, J. Marotzke, 2012: Forecast skill of multi-year seasonal means in the decadal prediction system of the Max Planck Institute for Meteorology. – Geophys. Res. Lett. **39**, L22707, DOI: 10.1029/2012GL053326.

Müller, W.A., H. Pohlmann, F. Sienz, D. Smith, 2014: Decadal climate predictions for the period 1901–2010 with a coupled climate model. – Geophys. Res. Lett. **41**, 2100–2107, DOI: 10.1002/2014GL059259.

Nissen, K.M., G.C. Leckebusch, J.G. Pinto, D. Renggli, S. Ulbrich, U. Ulbrich, 2010: Cyclones causing wind storms in the Mediterranean: characteristics, trends and links to large-scale patterns. – Nat. Hazard Earth Sys. **10**, 1379–1391, DOI: 10.5194/nhess-10-1379-2010.

Nissen, K., U. Ulbrich, G. Leckebusch, I. Kuhnel, 2014a: Decadal windstorm activity in the North Atlantic-European sector and its relationship to the meridional overturning circulation in an ensemble of simulations with a coupled climate model. – Climate Dyn. **43**, 1545–1555, DOI: 10.1007/s00382-013-1975-6.

Nissen, K.M., G.C. Leckebusch, J.G. Pinto, U. Ulbrich, 2014b: Mediterranean cyclones and windstorms in a changing climate. – Reg. Environ. Change **14**, 1873–1890, DOI: 10.1007/s10113-012-0400-8.

Pardowitz, T., D.J. Befort, G.C. Leckebusch, U. Ulbrich, submitted: Estimating uncertainties from high resolution simulations of extreme wind storms and consequences for impacts. – Meteor. Z.

Pohlmann, H., W.A. Müller, K. Kulkarni, M. Kameswarrao, D. Matei, F. Vamborg, C. Kadow, S. Illing, J. Marotzke, 2013a: Improved forecast skill in the tropics in the new MiKlip decadal climate predictions. – Geophys. Res. Lett. **40**, 5798–5802, DOI: 10.1002/2013GL058051.

Pohlmann, H., D.M. Smith, M.A. Balmaseda, N.S. Keenlyside, S. Masina, D. Matei, W.A. Müller, P. Rogel, 2013b: Predictability of the mid-latitude Atlantic meridional overturning circulation in a multi-model system. – Climate Dyn. **41**, 775–785, DOI: 10.1007/s00382-013-1663-6.

Polkova, I., A. Köhl, D. Stammer, 2014: Impact of initialization procedures on the predictive skill of a coupled ocean-atmosphere model. – Climate Dyn. **42**, 3151–3169, DOI: 10.1007/s00382-013-1969-4.

Renggli, D., G.C. Leckebusch, U. Ulbrich, S.N. Gleixner, E. Faust, 2011: The Skill of Seasonal Ensemble Prediction Systems to Forecast Wintertime Windstorm Frequency over the North Atlantic and Europe. – Mon. Wea. Rev. **139**, 3052–3068, DOI: 10.1175/2011MWR3518.1.

Robson, J.I., R.T. Sutton, D.M. Smith, 2012: Initialized decadal predictions of the rapid warming of the North Atlantic Ocean in the mid 1990s. – Geophys. Res. Lett. **39**, L19713, DOI: 10.1029/2012GL053370.

Robson, J., R. Sutton, D. Smith, 2014: Decadal predictions of the cooling and freshening of the North Atlantic in the 1960s and the role of ocean circulation. – Climate Dyn. **42**, 2353–2365, DOI: 10.1007/s00382-014-2115-7.

Scaife, A.A., M. Athanassiadou, M. Andrews, A. Arribas, M. Baldwin, N. Dunstone, J. Knight, C. MacLachlan, E. Manzini, W.A. Müller, H. Pohlmann, D. Smith, T. Stockdale, A. Williams, 2014: Predictability of the quasi-biennial oscillation and its northern winter teleconnection on seasonal to decadal timescales. – Geophys. Res. Lett. **41**, 1752–1758, DOI: 10.1002/2013GL059160.

Schwierz, C., P. Koellner-Heck, E.Z. Mutter, D.N. Bresch, P.-L. Vidale, M. Wild, C. Schaer, 2010: Modelling European winter wind storm losses in current and future climate. – Climatic Change **101**, 485–514, DOI: 10.1007/s10584-009-9712-1.

Sienz, F., W.A. Müller, H. Pohlmann, 2016: Ensemble size impact on the decadal predictive skill assessment. – Meteorol. Z. **25**, 645–655, DOI: 10.1127/metz/2016/0670.

Smith, D.M., R. Eade, N.J. Dunstone, D. Fereday, J.M. Murphy, H. Pohlmann, A.A. Scaife, 2010: Skilful multi-year predictions of Atlantic hurricane frequency. – Nat. Geosci. **3**, 846–849, DOI: 10.1038/NGEO1004.

Smith, D.M., R. Eade, H. Pohlmann, 2013: A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction. – Climate Dyn. **41**, 3325–3338, DOI: 10.1007/s00382-013-1683-2.

Solomon, A., L. Goddard, A. Kumar, J. Carton, C. Deser, I. Fukumori, A.M. Greene, G. Hegerl, B. Kirtman, Y. Kushnir, M. Newman, D. Smith, D. Vimont, T. Delworth, G.A. Meehl, T. Stockdale, 2011: Distinguishing the Roles of Natural and Anthropogenically Forced Decadal Climate Variability. – Bull. Amer. Meteor. Soc. **92**, 141–156, DOI: 10.1175/2010BAMS2962.1.

Stevens, B., M. Giorgetta, M. Esch, T. Mauritsen, T. Crueger, S. Rast, M. Salzmann, H. Schmidt, J. Bader, K. Block, R. Brokopf, I. Fast, S. Kinne, L. Kornblueh, U. Lohmann, R. Pincus, T. Reichler, E. Roeckner, 2013: Atmospheric component of the MPI-M Earth System Model: ECHAM6. – J. Adv. Model. Earth Syst. **5**, 146–172, DOI: 10.1002/jame.20015.

Stolzenberger, S., R. Glowienka-Hense, T. Spangehl, M. Schröder, A. Mazurkiewicz, A. Hense, 2015: Revealing skill of the MiKlip decadal prediction system by three dimensional probabilistic evaluation – Meteorol. Z. **25**, 657–671, DOI: 10.1127/metz/2015/0606.

Taylor, K.E., R.J. Stouffer, G.A. Meehl, 2012: An Overview of CMIP5 and the Experiment Design. – Bull. Amer. Meteor. Soc. **93**, 485–498, DOI: 10.1175/BAMS-D-11-00094.1.

Ulbrich, U., J.G. Pinto, H. Kupfer, G.C. Leckebusch, T. Spangehl, M. Reyers, 2008: Changing northern hemisphere storm tracks in an ensemble of IPCC climate change simulations. – J. Climate **21**, 1669–1679, DOI: 10.1175/2007JCLI1992.1.

Ulbrich, U., G.C. Leckebusch, J. Pinto, 2009: Extra-tropical cyclones in the present and future climate: a review. – Theor. Appl. Climatol. **96**, 117–131.

Uppala, S., P. Kållberg, A. Simmons, U. Andrae, V. da Costa Bechtold, M. Fiorino, J. Gibson, J. Haseler, A. Hernandez, G. Kelly, X. Li, K. Onogi, S. Saarinen, N. Sokka, R. Allan, E. Andersson, K. Arpe, M. Balmaseda, A. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, S. Caires, F. Chevallier, A. Dethof, M. Dragosavac, M. Fisher, M. Fuentes, S. Hagemann, E. Holm, B. Hoskins, L. Isaksen, P. Janssen, R. Jenne, A. McNally, J.-F. Mahfouf, J.-J. Morcrette, N. Rayner, R. Saunders, P. Simon, A. Sterl, K. Trenberth, A. Untch, D. Vasiljevic, P. Viterbo,, J. Woollen, 2005: The ERA-40 re-analysis. – Quart. J. Roy. Meteor. Soc. **131**, 2961–3012.

van Oldenborgh, G.J., F.J. Doblas-Reyes, B. Wouters, W. Hazeleger, 2012: Decadal prediction skill in a multi-model ensemble. – Climate Dyn. **38**, 1263–1280, DOI: 10.1007/s00382-012-1313-4.

Welker, C., O. Martius, 2014: Decadal-scale variability in hazardous winds in northern Switzerland since end of the 19th century. – Atmos. Sci. Let. **15**, 86–91, DOI: 10.1002/asl2.467.

Yeager, S., A. Karspeck, G. Danabasoglu, J. Tribbia, H. Teng, 2012: A Decadal Prediction Case Study: Late Twentieth-Century North Atlantic Ocean Heat Content. – J. Climate **25**, 5173–5189, DOI: 10.1175/JCLI-D-11-00595.1.

The pdf version (Adobe Java Script must be enabled) of this paper includes an electronic supplement:
Table of content – Electronic Supplementary Material (ESM)

Figures 10, 11, 12, 13, 14